GeoChat:Grounded Large Vision-Language Model for Remote Sensing

Year: 2023 | Citations: 273 | Authors: Kartik Kuckreja, M. S. Danish, Muzammal Naseer

---

Abstract

Recent advancements in Large Vision-Language Models (VLMs) have shown great promise in natural image domains, allowing users to hold a dialogue about given visual content. However, such general-domain VLMs perform poorly for Remote Sensing (RS) scenarios, leading to inaccurate or fabricated information when presented with RS domain-specific queries. Such a behavior emerges due to the unique challenges introduced by RS imagery. For example, to handle high-resolution RS imagery with diverse scale changes across categories and many small objects, region-level reasoning is necessary alongside holistic scene inter-pretation. Furthermore, the lack of domain-specific multimodal instruction following data as well as strong back-bone models for RS make it hard for the models to align their behavior with user queries. To address these limitations, we propose GeoChat - the first versatile remote sensing VLM that offers multitask conversational capabilities with high-resolution RS images. Specifically, GeoChat can not only answer image-level queries but also accepts region inputs to hold region-specific dialogue. Further-more, it can visually ground objects in its responses by referring to their spatial coordinates. To address the lack of domain-specific datasets, we generate a novel RS multimodal instruction-following dataset by extending image-text pairs from existing diverse RS datasets. We establish a comprehensive benchmarkfor RS multitask conversations and compare with a number of baseline methods. GeoChat demonstrates robust zero-shot performance on various RS tasks, e.g., image and region captioning, visual question answering, scene classification, visually grounded conversations and referring detection. Our code is available here.